Supplementary material for AAAI'19 Submission # 6931

Appendix

Proposition 3. *The variance returned by Algorithm 1 is*

$$\sigma^2_{T,entire}(\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{x}) - \sum_{i,j} \boldsymbol{z}_i^\top (\sum_l \boldsymbol{D}_l \boldsymbol{K}_{l,T} \boldsymbol{D}_l)^{-1} \boldsymbol{z}_j \quad (6)$$

where $\boldsymbol{D}_j = diag([g_j(\boldsymbol{x}_1), ..., g_j(\boldsymbol{x}_T)])$ *and* $\boldsymbol{z}_i = \boldsymbol{D}_i \boldsymbol{k}_{j,T}(\boldsymbol{x}) g_j(\boldsymbol{x}) \in \mathbb{R}^T$.

Proof of Proposition 3. According to Equation (4), the posterior covariance $k_{T,\text{entire}}(\boldsymbol{x},\boldsymbol{x}')$ in Algorithm 1 can be written as:

$$k_{T,\text{entire}}(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x},\boldsymbol{x}') - \boldsymbol{k}_T(\boldsymbol{x})^\top {\boldsymbol{K}_T}^{-1} \boldsymbol{k}_T(\boldsymbol{x}) \quad (11)$$

By the decomposition assumption (Equation (1)), we have $k(\boldsymbol{x},\boldsymbol{x}') = \sum_{j=1}^{J} g_j(\boldsymbol{x}) k_j(\boldsymbol{x},\boldsymbol{x}') g_j(\boldsymbol{x}')$. Moreover,

$$\begin{aligned}
&\boldsymbol{k}_T(\boldsymbol{x}) = [k(\boldsymbol{x}_1,\boldsymbol{x}), ..., k(\boldsymbol{x}_T,\boldsymbol{x})]^\top \\
&= \sum_{j=1}^{J} [g_j(\boldsymbol{x}_1) k_j(\boldsymbol{x}_1,\boldsymbol{x}) g_j(\boldsymbol{x}), ..., g_j(\boldsymbol{x}_T) k_j(\boldsymbol{x}_T,\boldsymbol{x}) g_j(\boldsymbol{x})]^\top \\
&= \sum_{j=1}^{J} \boldsymbol{D}_j \boldsymbol{k}_{j,T}(\boldsymbol{x}) g_j(\boldsymbol{x}) \qquad (12)
\end{aligned}$$

where $\boldsymbol{k}_{j,T}(\boldsymbol{x}) = [k_j(\boldsymbol{x}_1,\boldsymbol{x}), ..., k_j(\boldsymbol{x}_T,\boldsymbol{x})]^\top$.

The variance function $\sigma^2_T(\boldsymbol{x})$ is just the value of covariance function with $\boldsymbol{x}' = \boldsymbol{x}$. Therefore, combining Equation (11) and (12), the variance can be written as:

$$\begin{aligned}
&\sigma^2_{T,\text{entire}}(\boldsymbol{x}) = k_{T,\text{entire}}(\boldsymbol{x},\boldsymbol{x}) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}_T(\boldsymbol{x})^\top \boldsymbol{K}_T^{-1} \boldsymbol{k}_T(\boldsymbol{x}) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \sum_{i,j} g_i(\boldsymbol{x}) \boldsymbol{k}_{i,T}(\boldsymbol{x})^\top \boldsymbol{D}_i^\top \boldsymbol{K}_T^{-1} \boldsymbol{D}_j \boldsymbol{k}_{j,T}(\boldsymbol{x}) g_j(\boldsymbol{x}) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \sum_{i,j} \boldsymbol{z}_i^\top \boldsymbol{K}_T^{-1} \boldsymbol{z}_j \qquad (13) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \sum_{i,j} \boldsymbol{z}_i^\top (\sum_l \boldsymbol{D}_l \boldsymbol{K}_{l,T} \boldsymbol{D}_l)^{-1} \boldsymbol{z}_j \qquad (14)
\end{aligned}$$

with $\boldsymbol{z}_i = \boldsymbol{D}_i \boldsymbol{k}_{j,T} g_j(\boldsymbol{x}) \in \mathbb{R}^n$ and equation (13) to (14) is coming from:

$$\begin{aligned}
\boldsymbol{K}_T &= [k(\boldsymbol{x},\boldsymbol{x}')]_{\boldsymbol{x},\boldsymbol{x}' \in A_T} + \text{diag}([\sigma^2(\boldsymbol{x}_t)]_{t\in[T]}) \qquad (15) \\
&= \sum_{j=1}^{J} [g_j(\boldsymbol{x}) k_j(\boldsymbol{x},\boldsymbol{x}') g_j(\boldsymbol{x}')]_{\boldsymbol{x},\boldsymbol{x}' \in A_T} \\
&\qquad + \text{diag}([g_j^2(\boldsymbol{x}_t) \sigma_j^2(\boldsymbol{x}_t)]_{t\in[T]}) \qquad (16) \\
&= \sum_j \boldsymbol{D}_j ([k_j(\boldsymbol{x},\boldsymbol{x}')]_{\boldsymbol{x},\boldsymbol{x}' \in A_T} + \text{diag}([\sigma_j^2(\boldsymbol{x}_t)]_{t\in[T]})) \boldsymbol{D}_j \\
&= \sum_j \boldsymbol{D}_j \boldsymbol{K}_{j,T} \boldsymbol{D}_j
\end{aligned}$$

where the first kernel term from Equation (15) to (16) is derived by from definition. And in the latter term, from the decomposition assumption (1), the noise variance $\sigma^2(\boldsymbol{x})$ of the target function f at point $\boldsymbol{x}$ is the cumulative variance of the noise variance $\sigma_j^2(\boldsymbol{x})$ of each individual function f_j, i.e.

$$\sigma^2(\boldsymbol{x}) = \sum_{j=1}^{J} g_j^2(\boldsymbol{x}) \sigma_j^2(\boldsymbol{x}) \; \forall \boldsymbol{x} \in \mathcal{X}, j \in [J]$$

which explains the derivation from Equation 15 to 16. □

Proposition 4. *The variance returned by Algorithm 2 is*

$$\sigma^2_{T,decomp}(\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{x}) - \sum\nolimits_l \boldsymbol{z}_l^\top (\boldsymbol{D}_l \boldsymbol{K}_{l,T} \boldsymbol{D}_l)^{-1} \boldsymbol{z}_l \quad (7)$$

Proof of Proposition 4. In Algorithm 2, it runs GP regression to each function $f_j(\boldsymbol{x})$ respectively. We can compute the corresponding posterior covariance function $k_{j,T}$ by:

$$k_{j,T}(\boldsymbol{x},\boldsymbol{x}') = k_j(\boldsymbol{x},\boldsymbol{x}') - \boldsymbol{k}_{j,T}(\boldsymbol{x})^\top \boldsymbol{K}_{j,T}^{-1} \boldsymbol{k}_{j,T}(\boldsymbol{x})$$

By Algorithm 2, the synthetic covariance of the target function $f(\boldsymbol{x})$ is:

$$k_{T,\text{decomp}}(\boldsymbol{x},\boldsymbol{x}') = \sum_{j=1}^{J} g_j(\boldsymbol{x}) k_{j,T}(\boldsymbol{x},\boldsymbol{x}') g_j(\boldsymbol{x}')$$

$$\begin{aligned}
&\sigma^2_{T,\text{decomp}}(\boldsymbol{x}) = k_{T,\text{decomp}}(\boldsymbol{x},\boldsymbol{x}) \\
&= \sum_{j=1}^{J} g_j(\boldsymbol{x}) k_{j,T}(\boldsymbol{x},\boldsymbol{x}) g_j(\boldsymbol{x}) \\
&= \sum_{j=1}^{J} g_j(\boldsymbol{x}) k_j(\boldsymbol{x},\boldsymbol{x}) g_j(\boldsymbol{x}) \\
&\qquad - \sum_{j=1}^{J} g_j(\boldsymbol{x}) \boldsymbol{k}_{j,T}(\boldsymbol{x})^\top \boldsymbol{K}_{j,T}^{-1} \boldsymbol{k}_{j,T}(\boldsymbol{x}) g_j(\boldsymbol{x}) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \sum_j \boldsymbol{z}_j^\top \boldsymbol{D}_j^{-1} \boldsymbol{K}_{j,T}^{-1} \boldsymbol{D}_j^{-1} \boldsymbol{z}_j \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \sum_j \boldsymbol{z}_j^\top (\boldsymbol{D}_j \boldsymbol{K}_{j,T} \boldsymbol{D}_j)^{-1} \boldsymbol{z}_j
\end{aligned}$$

□

Proof of Theorem 6

Theorem 6. *The variance provided by decomposed Gaussian process regression (Algorithm 2) is less than or equal to the variance provided by Gaussian process regression (Algorithm 1), which implies the uncertainty by using decomposed Gaussian process regression is smaller.*

Proof. If we write $\boldsymbol{B}_l = \boldsymbol{D}_l \boldsymbol{K}_{l,T} \boldsymbol{D}_l$, $\boldsymbol{B}$ is positive definite since it is the multiplication of positive definite matrix $\boldsymbol{K}_{l,T}$ and two $\boldsymbol{D}_l$ identical diagonal matrices. Using Proposition 3 and 4, the difference between Equation (6) and (7) can be written as:

$$\begin{aligned}
&\sigma^2_{T,\text{entire}}(\boldsymbol{x}) - \sigma^2_{T,\text{decomp}}(\boldsymbol{x}) \\
&= \sum_l \boldsymbol{z}_l^\top (\boldsymbol{D}_l \boldsymbol{K}_{l,T} \boldsymbol{D}_l)^{-1} \boldsymbol{z}_l - \sum_{i,j} \boldsymbol{z}_i^\top (\sum_l \boldsymbol{D}_l \boldsymbol{K}_{l,T} \boldsymbol{D}_l)^{-1} \boldsymbol{z}_j \\
&= \sum\nolimits_l \boldsymbol{z}_l^\top \boldsymbol{B}_l^{-1} \boldsymbol{z}_l - \sum\nolimits_{i,j} \boldsymbol{z}_i^\top (\sum\nolimits_l \boldsymbol{B}_l)^{-1} \boldsymbol{z}_j \\
&= \sum\nolimits_l \boldsymbol{z}_l^\top \boldsymbol{B}_l^{-1} \boldsymbol{z}_l - J(\frac{\sum_i \boldsymbol{z}_i}{J})^\top (\frac{\sum_l \boldsymbol{B}_l}{J})^{-1} (\frac{\sum_i \boldsymbol{z}_i}{J}) \\
&= \sum\nolimits_l h(\boldsymbol{B}_l, \boldsymbol{z}_l) - J h(\bar{\boldsymbol{B}}, \bar{\boldsymbol{z}}) \geq 0
\end{aligned}$$

where $\bar{\boldsymbol{B}} = \frac{\sum_i \boldsymbol{B}_i}{J}$ and $\bar{\boldsymbol{z}} = \frac{\sum_i \boldsymbol{z}_i}{J}$ are the average value. The last inequality comes from Jensen inequality and Lemma 5, which says the matrix-fractional function h is convex. □

Proof of Theorem 7

In order to prove this, we follow the similar techniques as GPUCB (Srinivas et al. 2009), which is illustrated as follows:

Lemma 9 (Modified version of Lemma 5.1 from Srinivas et al.)**.** *Given $f(\boldsymbol{x}) = \sum_{j=1}^J g_j(\boldsymbol{x}) f_j(\boldsymbol{x})$ (Definition 1), deterministic known functions g_j and unknown $f_j \sim GP(0, k_j(\boldsymbol{x}, \boldsymbol{x}'))$, pick $\delta \in (0,1)$ and set $\beta_t = 2\log(|\mathcal{X}|\pi_t/\delta)$, where $\sum_{t \geq 1} \pi_t^{-1} = 1, \pi_t > 0$. Then, the $\mu_{t-1}(\boldsymbol{x}), \sigma_{t-1}(\boldsymbol{x})$ returned by Algorithm 3 satisfy:*

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}) \forall \boldsymbol{x} \in \mathcal{X}, t \geq 1$$

with probability $1 - \delta$.

Proof. Fix $t \geq 1$ and $\boldsymbol{x} \in \mathcal{X}$, Conditioned on sampled points $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}\}$ and sampled values $\{y_{1,j}, ..., y_{t-1,j} \forall j \in [J]\}$, the Bayesian property of decomposed GP regression (Algorithm 2) implies that the function value at point $\boldsymbol{x}$ forms a Gaussian distribution with mean $\mu_{t-1}(\boldsymbol{x})$ and variance $\sigma^2_{t-1}(\boldsymbol{x})$, i.e. $f(\boldsymbol{x}) \sim N(\mu_{t-1}(\boldsymbol{x}), \sigma^2_{t-1}(\boldsymbol{x}))$. Now, if $r \sim N(0,1)$, then

$$\begin{aligned}
Pr\{r > c\} &= e^{-c^2/2}(2\pi)^{-1/2} \int e^{-(r-c)^2/2 - c(r-c)} dr \\
&\leq e^{-c^2/2} Pr\{r > 0\} = (1/2) e^{-c^2/2} \qquad (17)
\end{aligned}$$

for $c > 0$, since $e^{-c(r-c)} \leq 1$ for $r \geq c$. Therefore, $Pr\{|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x})\} \leq e^{-\beta_t^{1/2}}$, using $r = (f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x}))/\sigma_{t-1}(\boldsymbol{x})$ and $c = \beta_t^{1/2}$. Then apply the union bound to all $\boldsymbol{x} \in \mathcal{X}$:

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}) \; \forall \boldsymbol{x} \in \mathcal{X}$$

holds with probability $\geq 1 - |\mathcal{X}| e^{-\beta_t^{1/2}}$. Choosing $|\mathcal{X}| e^{-\beta_t^{1/2}} = \delta/\pi_t$ and using the union bound for $t \in \mathbb{N}$, the statement holds. For example, we can use $\pi_t = \pi^2 t^2/6$. □

The proof is almost the same as Theorem 5.1 in (Srinivas et al. 2009) except the Bayesian property of decomposed Gaussian process, where the Bayesian property of decomposed Gaussian process can be gotten from the Bayesian property of each individual function f_j and the linear combination of Gaussian distributions is still a Gaussian distribution, which implies the posterior belief after performing decomposed GP regression at a given point $\boldsymbol{x}$ still form a Gaussian distribution with composed mean and variance.

Lemma 10 (Modified version of Lemma 5.2 from Srinivas et al.)**.** *Fix $t \geq 1$. If $|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, then the regret r_t is bounded by $2\beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}_t)$, where $\boldsymbol{x}_t$ is the t-th choice of Algorithm 3.*

Proof. By definition of $\boldsymbol{x}_t$: $\mu_{t-1}(\boldsymbol{x}_t) + \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}_t) \geq \mu_{t-1}(\boldsymbol{x}^*) + \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^*)$. Therefore,

$$\begin{aligned}
r_t &= f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq \beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}_t) + \mu_{t-1}(\boldsymbol{x}_t) - f(\boldsymbol{x}_t) \\
&\leq 2\beta_t^{1/2} \sigma_{t-1}(\boldsymbol{x}_t)
\end{aligned}$$

□

Lemma 11 (Modified version of Lemma 5.3 from Srinivas et al.)**.** *The information gain for the points selected can be expressed in terms of the predictive variances. If $\boldsymbol{f_{j,T}} = \{f_j(\boldsymbol{x}_t)\}_{t \in [T]} \in \mathbb{R}^T$ and $\boldsymbol{y_{j,T}} = \{y_{j,t}\}_{t \in [T]} \in \mathbb{R}^T$:*

$$I(\boldsymbol{y_{j,T}} : \boldsymbol{f_{j,T}}) = \frac{1}{2} \sum\nolimits_{t=1}^T \log(1 + \sigma^{-2} \sigma^2_{j,t-1}(\boldsymbol{x}_t))$$

where $f_j(\boldsymbol{x}_t), y_{j,t}, \sigma^2_{j,t-1}$ follow the definition and derivation in Algorithm 3.

Proof. Directly follow by replacing all the f, y, σ by f_j, y_j, σ_j in the proof of Theorem 5.3 from Srinivas et al.. □

Theorem 7. *Let $\delta \in (0,1)$ and $\beta_t = 2\log(|\mathcal{X}| t^2 \pi^2/6\delta)$. Running decomposed GP-UCB (Algorithm 3) for a composed sample $f(\boldsymbol{x}) = \sum_{j=1} g_j(\boldsymbol{x}) f_j(\boldsymbol{x})$ with bounded variance $k_j(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ and each $f_j \sim GP(0, k_j(\boldsymbol{x}, \boldsymbol{x}'))$, we obtain a regret bound of $\mathcal{O}(\sqrt{T \log|\mathcal{X}| \sum_{j=1}^J B_j^2 \gamma_{j,T}})$ with high probability, where $B_j = \max_{\boldsymbol{x} \in \mathcal{X}} |g_j(\boldsymbol{x})|$. Precisely,*

$$Pr\Big\{R_T \leq \sqrt{C_1 T \beta_T \sum\nolimits_{j=1}^J B_j^2 \gamma_{j,T}} \; \forall T \geq 1\Big\} \geq 1 - \delta \quad (8)$$

where $C_1 = 8/\log(1 + \sigma^{-2})$ with noise variance σ^2.

Proof. According to Lemma 11, we can take advantage of the individual information gain of each $f_j(\boldsymbol{x})$, which is

$$\begin{aligned}
I_j(y_{j,T}; f_j) &= \frac{1}{2} \sum_{t=1}^T \log(1 + \sigma^{-2} \sigma^2_{j,t-1}(\boldsymbol{x}_t)) \\
\gamma_{j,T} &= \max I_j(y_{j,T}; f_j)
\end{aligned}$$

Besides, we can also bound the total regret by the individual information gains as following:

$$\begin{aligned}\sum_{j=1}^{J} B_j^2 I_j(y_{j,T}; f_j) &= \frac{1}{2}\sum_{j=1}^{J} B_j^2 \sum_{t=1}^{T} \log(1+\sigma^{-2}\sigma_{j,t-1}^2(\boldsymbol{x}_t)) \\ &\geq \frac{1}{2}\sum_{j=1}^{J} B_j^2 \sum_{t=1}^{T} C_2^{-1}\sigma^{-2}\sigma_{j,t-1}^2(\boldsymbol{x}_t) \\ &\geq \frac{1}{2}C_2^{-1}\sigma^{-2}\sum_{j=1}^{J}\sum_{t=1}^{T} g_j^2(\boldsymbol{x}_t)\sigma_{j,t-1}^2(\boldsymbol{x}_t) \\ &\geq \frac{1}{2}C_2^{-1}\sigma^{-2}\sum_{t=1}^{T}\frac{r_t^2}{4\beta_t} \\ &\geq \frac{C_2^{-1}\sigma^{-2}}{8\beta_T}\sum_{t=1}^{T} r_t^2\end{aligned}$$

where $C_2 = \sigma^{-2}/\log(1+\sigma^{-2}) \geq 1$, $s^2 \leq C_2\log(1+s^2)$ for $s \in [0, \sigma^{-2}]$ and $\sigma^{-2}\sigma_{j,t-1}^2(\boldsymbol{x}_t) \leq \sigma^{-2}k_j(\boldsymbol{x}_t, \boldsymbol{x}_t) \leq \sigma^{-2}$. Let $C_1 = 8\sigma^2 C_2 = 8/\log(1+\sigma^{-2})$. Applying Cauchy inequality gives us:

$$C_1\beta_T T\sum_{j=1}^{J} B_j^2 I_j(y_{j,T}; f_j) \geq (\sum_{t=1}^{T} r_t)^2 = R_T^2$$

which implies a similar upper bound

$$R_T \leq \sqrt{C_1 T\beta_T \sum_{j=1}^{J} B_j^2\gamma_{j,T}}$$

□

Proof of Theorem 8

All the proofs in Theorem 7 apply except Lemma 9. Since the decomposition here is non-linear, therefore the composition of outcomes of Gaussian processes is no longer an outcome of Gaussian process, which prohibits us to have a nice Gaussian process property: function value $f(\boldsymbol{x})$ does not form a Gaussian distribution. Due to the non-linearity, the distribution gets distorted, losing its original form with Gaussian distribution. Fortunately, if the partial derivatives of function $g : \mathbb{R}^J \to \mathbb{R}$ (Definition 2) are bounded, then we can still perform a similar estimation and bound the distribution by a larger Gaussian distribution, which enables us to have a similar result.

Lemma 12 (General Version with Definition 2 and Algorithm 4)**.** *Given $f(\boldsymbol{x}) = g(f_1(\boldsymbol{x}), ..., f_J(\boldsymbol{x}))$ (Definition 2), deterministic known functions g and unknown $f_j \sim GP(0, k_j(\boldsymbol{x}, \boldsymbol{x}'))$, pick $\delta \in (0,1)$ and set $\beta_t = 2\log(|\mathcal{X}|J\pi_t/\delta)$, where $\sum_{t\geq 1}\pi_t^{-1} = 1, \pi_t > 0$. Further assume the function g has bounded partial derivatives $B_j = \max_{\boldsymbol{x}\in\mathcal{X}}|\nabla_j g(\boldsymbol{x})|\ \forall j \in [J]$. Then, the $\mu_{t-1}(\boldsymbol{x}), \sigma_{t-1}(\boldsymbol{x})$ returned by Algorithm 4 satisfy:*

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}) \forall \boldsymbol{x} \in \mathcal{X}, t \geq 1$$

with probability $1-\delta$.

Proof. The main problem here is the posterior distribution of $f(\boldsymbol{x})$ is not a Gaussian distribution. But fortunately, the posterior distribution of each $f_j(\boldsymbol{x})$ is still a Gaussian distribution with mean $\mu_{j,t-1}(\boldsymbol{x})$ and variance $\sigma_{j,t-1}^2(\boldsymbol{x})$ for any given $\boldsymbol{x} \in \mathcal{X}$. Then,

$$\begin{aligned}&|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \\ &= |g(f_1(\boldsymbol{x}), ..., f_J(\boldsymbol{x})) - g(\mu_{1,t-1}(\boldsymbol{x}), ..., \mu_{J,t-1}(\boldsymbol{x}))| \\ &\leq \sum\nolimits_{j=1}^{J} B_j|f_j(\boldsymbol{x}) - \mu_{j,t-1}(\boldsymbol{x})| \qquad (18)\end{aligned}$$

Applying the same argument in Lemma 9 to function f_j:

$$Pr\{|f_j(\boldsymbol{x}) - \mu_{j,t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{j,t-1}(\boldsymbol{x})\} \leq e^{-\beta_t^{1/2}}$$

Then applying the union bound on $j \in [J]$, we get

$$\begin{aligned}&|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \\ &\leq \sum\nolimits_{j=1}^{J} B_j|f_j(\boldsymbol{x}) - \mu_{j,t-1}(\boldsymbol{x})| \\ &\leq \sum\nolimits_{j=1}^{J} B_j\beta_t^{1/2}\sigma_{j,t-1}(\boldsymbol{x}) \\ &\leq \beta_t^{1/2}\sqrt{J(\sum\nolimits_{j=1}^{J} B_j^2\sigma_{j,t-1}^2(\boldsymbol{x}))} \\ &= \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})\end{aligned}$$

with probability $1 - Je^{-\beta_t^{1/2}}$, where the last inequality is from Cauchy's inequality. Then apply union bound again to all $\boldsymbol{x} \in \mathcal{X}$, the above inequality yields:

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathcal{X}$$

with probability $1 - |\mathcal{X}|Je^{-\beta_t^{1/2}}$. Choosing $|\mathcal{X}|Je^{-\beta_t^{1/2}} = \delta/\pi_t$ and using the union bound for $t \in \mathbb{N}$, the statement holds, i.e. $\beta_t = 2\log(|\mathcal{X}|J\pi_t/\delta)$. Specifically, if we choose $\pi_t = \pi^2 t^2/6$, then it implies $\beta_t = 2\log(|\mathcal{X}|Jt^2\pi^2/6\delta)$. □

Theorem 8. *By running generalized decomposed GP-UCB with $\beta_t = 2\log(|\mathcal{X}|Jt^2\pi^2/6\delta)$ for a composed sample $f(\boldsymbol{x}) = g(f_1(\boldsymbol{x}), ..., f_J(\boldsymbol{x}))$ of GPs with bounded variance $k_j(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ and each $f_j \sim GP(0, k_j(\boldsymbol{x}, \boldsymbol{x}'))$. we obtain a regret bound of $\mathcal{O}(\sqrt{T\log|\mathcal{X}|\sum_{j=1}^{J} B_j^2\gamma_{j,T}})$ with high probability, where $B_j = \max_{\boldsymbol{x}\in\mathcal{X}}|\nabla_j g(\boldsymbol{x})|$. Precisely,*

$$Pr\big\{R_T \leq \sqrt{C_1 T\beta_T\sum\nolimits_{j=1}^{J} B_j^2\gamma_{j,T}}\ \forall T \geq 1\big\} \geq 1-\delta \quad (10)$$

where $C_1 = 8/\log(1+\sigma^{-2})$ with noise variance σ^2.

Proof. Directly follow by the same proofs of Theorem 7 with Lemma 10, Lemma 11, and Lemma 12. □

Remark 1. *In inequality* (18)*, if we write $Z_j = |f_j(\boldsymbol{x}) - \mu_{j,t-1}(\boldsymbol{x})|$, where $f_j(\boldsymbol{x}) - \mu_{j,t-1}(\boldsymbol{x})$ is sampled from a normal distribution with 0 mean and $\sigma_{j,t-1}^2(\boldsymbol{x})$ (due to Gaussian process property). Then this Z_j is a random variable drawn from a half-normal distribution with parameter $\sigma_j(\boldsymbol{x})$ (no longer the variance here).*

The summation of half-normal distributions can still be computed and bounded by a similar inequality like inequality (17)*. This can provide a constant ratio of improvement to the β_t exploration parameter, thus the regret bound as well. However it does not change the order of regret and sample complexity. Therefore we are not going to cover this here.*